

On the use and interpretation of Generalized Linear Models (GLMs) in agricultural sciences: critical analysis and application guidelines

Bruno Enagnon Lokonon ^(1,*), **Paulette Marthe Behanzin Guedezeoume** ⁽¹⁾,
Scholastique M. Merveille Essetcheou ⁽¹⁾ and **Romain Glèlè Kakai** ⁽¹⁾

⁽¹⁾ Laboratoire de Biomathématiques et d'Estimations Forestières (LABEF), Faculté des Sciences Agronomiques, Université d'Abomey-Calavi, 04 BP 1525, Cotonou, Bénin

Received on September 18, 2023; Accepted on December 23, 2028; Published on December 31, 2023

Copyright © 2023, The African Journal of Applied Statistics (AJAS) and The Statistics and Probability African Society (SPAS). All rights reserved

Abstract. Generalized Linear Models (GLMs) provide a unified approach that encompasses most of the linear models. However, they could not been properly used in agricultural sciences. In this review, we (1) examined the correctness related to the application of GLMs to agricultural sciences; and (2) provided a guideline for their suitable use. From a total of 832 articles identified, 155 articles were retained for the review. The results revealed that the use of GLMs in agricultural sciences has increased over time. However, in most cases (57.42%), GLMs outputs were not well interpreted. Moreover, useful expected findings from the use of GLMs were not reported in several papers. Indeed, 27.95% of the selected articles did not assess the goodness of the models and 21.30% did not report the significance of the coefficients. Our work has revealed the gap between the introduction of a statistical method and effective use in real world situations. A practical guideline with regard to GLMs has been provided.

Key words: Generalized Linear Models; non-normal error structures; effective use; agronomy; critical review.

AMS 2010 Mathematics Subject Classification Objects : 62J12; 92B05 ; 92B15.

(*) Corresponding author: Bruno Enagnon Lokonon (brunolokonon@gmail.com)
Paulette Marthe Behanzin Guedezeoume: pguedezeoume5@gmail.com;
Scholastique M. Merveille Essetcheou: 2meyoesset@gmail.com

Résumé (Abstract in French) Les modèles linéaires généralisés (GLMs) constituent une approche unifiée qui englobe la plupart des modèles linéaires. Cependant, ils pourraient être mal utilisés dans les sciences agricoles. Dans cette revue, nous avons (1) examiné l'exactitude de l'application des GLMs aux sciences agricoles et (2) fourni une ligne directrice pour leur utilisation appropriée. Sur un total de 832 articles identifiés, 155 ont été retenus pour l'examen. Les résultats ont révélé que l'utilisation des GLMs dans les sciences agricoles a augmenté au fil du temps. Toutefois, dans la plupart des cas (57,42 %), les résultats des GLMs n'ont pas été bien interprétés. En outre, les résultats utiles attendus de l'utilisation des GLMs n'ont pas été rapportés dans plusieurs articles. Par exemple, 27,95 % des articles sélectionnés n'ont pas évalué la qualité des modèles. Notre travail a révélé l'écart entre l'introduction d'une méthode statistique et son utilisation effective dans des situations réelles. Un guide pratique concernant les GLMs a été fourni.

Presentation of authors.

Bruno Enagnon Lokonon, PhD, is a Research fellow in Biometry and ecology, affiliated to: *Laboratoire de Biomathématiques et d'Estimations Forestières (LABEF)*, University of Abomey-Calavi (UAC), Benin

Paulette Marthe Béhanzin Guedezeoumé, M Sc, is a PhD student in Biometry, at: *Laboratoire de Biomathématiques et d'Estimations Forestières (LABEF)*, University of Abomey-Calavi (UAC), Benin

Scholastique M. Merveille Essetcheou, MSc. in Biometry, is affiliated to: *Laboratoire de Biomathématiques et d'Estimations Forestières (LABEF)*, University of Abomey-Calavi (UAC), Benin

Romain Glèlè Kakaï, Ph.D., is Professor of Biometrics and Forest estimations. He is the Head of the *Laboratoire de Biomathématiques et d'Estimations Forestières (LABEF)*, the Coordinator of the doctoral studies in Biometry and the President of the Scientific Council of Agronomic Sciences, University of Abomey-Calavi (UAC) Benin

1. Introduction

In several research fields such as agriculture, economy, and sociology, researchers are interested in modelling the relationship between variables. In many cases, the linear models are sufficient to correctly describe the link between a response variable and one or more explanatory variables, and to explain a biological phenomenon (Bel *et al.*, 2016; Ruíz *et al.*, 2023). However, the use of the linear models require some assumptions to be validated: the errors are independent, normally distributed, with a mean of 0 and constant variance (Neter *et al.*, 1996; Ruíz *et al.*, 2023). In most situations, real data do not satisfy these assumptions, limiting the use of linear models (Agresti, 2015). Generalized Linear Models (GLMs) are an extension of the classical linear models allowing to overcome these assumptions and to model observations whose probability law belongs to the exponential family (Ruíz *et al.*, 2023). First presented under this name by Nelder and Wedderburn (1972), GLMs have been deeply discussed by McCullagh and Nelder (1989).

GLMs have several areas of application ranging from medicine to conservation biology. For instance, in epidemiology, the logistic model is widely used to provide a basic tool for epidemiological investigation of chronic diseases (Khuri *et al.*,

2006). In entomology, *GLMs* are applied to relate changes in insect behavior to changes in the chemical composition of a plant extracts (Hern and Dorn, 2001). In conservation biology, presence/absence or count data are most analyzed using *GLMs* with an appropriate error structures (Bourne *et al.*, 2007).

However, despite its flexibility in dealing with complex responses and many variables, *GLMs* are challenging for applied researchers (Pekár and Brabec, 2016). One the other hand, in agricultural sciences and related fields, for instance, non normal data are common (Bolker *et al.*, 2009; O'Hara and Kotze, 2010). Examples include presence/absence of species in plots (binomial), weed count or number of parasites per plot (Poisson or negative binomial), time to flowering (exponential or gamma), disease rating category (multinomial), and proportion of leaf area affected (beta) (Bolker *et al.*, 2009; Stroup, 2015). This can explain the increasing number of publications in agricultural sciences that use *GLMs* (Bourne *et al.*, 2007). However, the effective use of this statistical method in the field of agricultural research is not always assessed. Indeed, researchers with basic background in statistics/biostatistics are limited in their ability to use *GLMs* to analyze their data (Bourne *et al.*, 2007). This often leads to incorrect use and misinterpretation of the results (Pekár and Brabec, 2016). Researchers can erroneously conclude that variables have a meaningful effect when in fact they do not (Zuur *et al.*, 2009). The necessity for accurate biological inference therefore calls to adequately employ *GLMs* to minimize the risk of Type I error (Hilbe, 2011). Moreover, the purpose of data analysis is decision making. Therefore, the misuse of *GLMs* can lead to a wrong decision and can also make the research findings difficult to replicate (Kim *et al.*, 2011). The misuse and misinterpretation of statistical methods may have serious and irreversible consequences such as loss of life.

The main objective of this study is to contribute to an effective use of *GLMs* in agricultural sciences. The research question addressed is: are the *GLMs* effectively applied in agricultural sciences and related fields for data analysis? Specifically, the study has (1) reviewed the use of *GLMs* for data analysis in agricultural sciences and related fields; (2) identified research and knowledge gaps from the use of *GLMs* in agricultural sciences and related fields; and (3) proposed guidelines for effective use of *GLMs* and provided an example which is apply to agricultural data.

The rest of the paper is organized as follows. In Section 1, we present the problematic linked to the use of *GLMs* and the objectives of the present study. Section 2 presents the principles of *GLMs*, while Section 3 describes the methodological aspects of the study. The main findings from the study are presented in Section 4. Section 5 formulates the research and knowledge gaps. Sections 6 and 7 respectively present the general guidelines for an effective use of *GLMs* and the concluding remarks. In Appendix, an application guideline in **R** software is presented.

2. Principle of *GLMs*

GLMs are extensions of the classical linear models (Guisan *et al.*, 2002). A linear model describing the relationship between $Y = (y_1, \dots, y_n)$, a vector of observations (a continuous response variable) and X , the matrix of explanatory variables, is written as:

$$Y = X\beta + \epsilon \quad (1)$$

where β and ϵ represent the vectors of model parameters and residuals, respectively.

It is also possible to write the same model through the expectancy of Y considering X without explicitly specifying ϵ .

$$E(Y|X) = \mu = X\beta \quad (2)$$

This linear model can be transformed to a generalized linear model (Olsson, 2015) by replacing μ by $g(\mu)$ in equation (2).

$$g(\mu) = X\beta = \eta \quad (3)$$

η is a linear combination of the covariates and $g(\cdot)$ defines the relationship between μ and η . $g(\cdot)$ is called a link function.

GLMs is defined through three components:

1. a random component which specifies the distribution of the response variable belonging to the exponential family, whose probability density function is:

$$f(y; \theta; \varphi) = \exp \left\{ \frac{y\theta - b(\theta)}{a(\varphi)} + c(y; \varphi) \right\} \quad (4)$$

where $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are known functions, φ is a dispersion parameter and θ a canonical parameter (McCullagh and Nelder, 1989). Many commonly used distributions in the exponential family are the Normal, Binomial, Poisson, and Gamma distributions.

2. a systematic component, which defines a linear combination of the covariates in the model, such that $\eta = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$.
3. a link function $g(\cdot)$ relating η to μ so that $\eta = g(\mu)$. If $g(\cdot)$ is the identity function, we obtain the special class of ordinary linear models and when $g(\cdot)$ is the $\log$ function, we are in the case of Poisson models.

A GLM has the following properties: $\theta = g(\mu)$, $E(Y) = \mu = b'(\theta)$ and $Var(Y) = a(\varphi)b''(\theta)$. For instance, Poisson distribution can be written as a special case of an exponential family distribution as follows:

$$P[Y = y; \lambda] = \frac{\lambda^y}{y!} e^{-\lambda} = \exp(y \log \lambda - \lambda - \log(y!)) \quad (5)$$

By Comparing the equation (5) with (4), it is shown that (5) is a special case of (4) with $\theta = \log(\lambda)$, $b(\theta) = \lambda$, $c(y, \varphi) = -\log(y!)$ and $a(\varphi) = 1$. Poisson distribution is therefore a special case of exponential family distribution and is the random component of Poisson model.

In GLMs, parameters estimation is carried out through the maximization of the likelihood function. The likelihood function in y is:

$$\begin{aligned} \mathcal{L}(y; \theta, \varphi) &= \prod_{i=1}^n f(y_i; \theta_i, \varphi) \\ &= \prod_{i=1}^n f(y_i; x_i \beta, \varphi) \\ &= \prod_{i=1}^n \exp \left\{ \frac{y_i x_i \beta - b(x_i \beta)}{a(\varphi)} + c(y_i, \varphi) \right\} \end{aligned} \quad (6)$$

The log-likelihood becomes:

$$\ell(y; \beta, \varphi) = \sum_{i=1}^n \left[\frac{1}{a(\varphi)} (y_i x_i \beta - b(x_i \beta)) + c(y_i, \varphi) \right] \quad (7)$$

To obtain the maximum likelihood estimates, iterative optimization algorithms are used. The most used is the Newton-Raphson algorithm (Bel *et al.*, 2016) whose expression at the $(k + 1)$ -th iteration gives:

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} - \mathbf{H}^{-1} \left(\hat{\beta}^{(k)} \right) \frac{\partial \ell(y, \beta)}{\partial \beta} \quad (8)$$

where $\mathbf{H} \left(\hat{\beta}^{(k)} \right) = \frac{\partial^2 \ell(y, \beta)}{\partial \beta \partial \beta^T}$ is the Hessian matrix.

3. Methods

The review was conducted following Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Figure 1). Published papers in agricultural sciences and related fields were identified using platforms such as MDPI, Science Direct and Google Scholar. The keywords used in the search were: "Generalized Linear Models", "GLMs", "use of GLMs in agricultural sciences". The following related fields were included separately in the search in combination with the keywords: "agronomy", "food nutrition", and "ecology". The review was conducted considering the period from 2009 to 2022. In total, 832 papers were recorded. Commentaries, purely descriptive studies, and letters were excluded. The exclusion criteria were chosen to help in the elimination of studies that are outside the scope of this study. This resulted in 155 selected papers, which were retained for the review.

From each selected paper, we recorded the year, field, type of data, whether GLMs were required to analyse data regarding the methodology and the study's objectives or not, and whether GLMs were used or not, and the distribution of the response variable. We checked if the goodness of the model was fitted, and which kind of goodness fit test was used, and if the significance test is done and interpreted. Moreover, we recorded the software used to perform the analysis and finally the suitable interpretation of outputs.

Count and relative frequencies were computed for the recorded data and tables and figures were drawn to summarize the information. Research and knowledge gaps have been identified, guidelines suggested and application to real data discussed.

4. Assessment of effective use of GLMs in agricultural sciences and related fields

Four research fields were recorded from the selected articles, namely crop production, genetics, ecology and evolution and food nutrition (Table 1). The research field with the highest number of articles was food nutrition (37.42%) followed by ecology and evolution (17.42%). Genetics (3.23%) had the smallest number of the reviewed papers considered.

The use of GLMs in agricultural sciences and related fields increased over time (Figure 2). From 2009 to 2022, the use of GLMs in agricultural sciences and related fields has increased five-fold (Figure 2).

Two distributions of exponential family were mostly used in the selected papers, namely Binomial (41.94%) and Poisson (21.94%) distributions (Figure 3). Other distributions such as Negative Binomial (11.61%) and Gamma (1.29%) were also encountered. Moreover, 19.35% of the selected papers did not report the response variable (Figure 3). Reporting the distribution of the response variable is important since different link functions may be used for a given distribution. For instance,

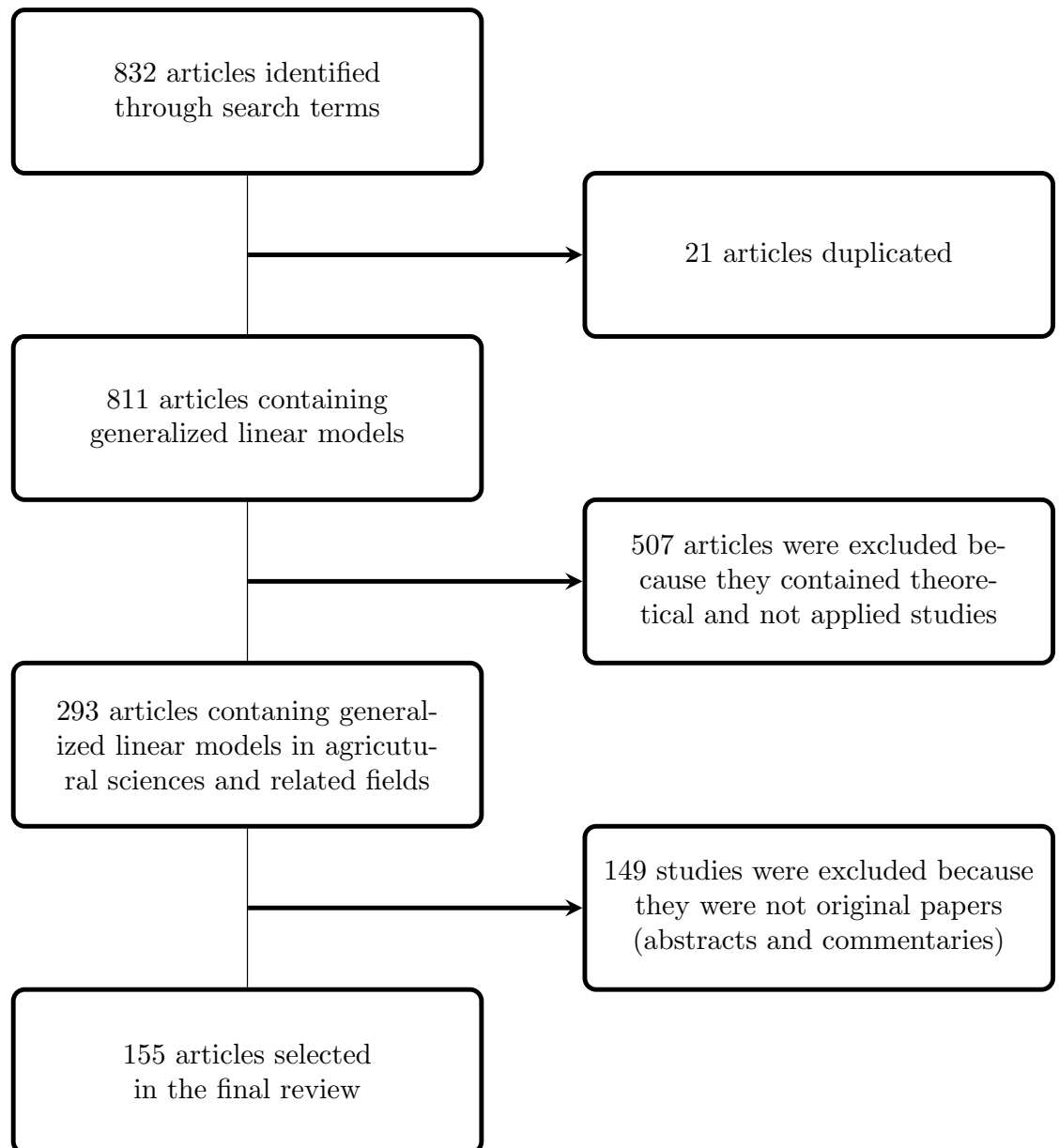


Fig. 1: Flow chart diagram of papers selection process

binary data may be suitably analyzed with probit link function instead of logit link function.

In the selected papers, four goodness of fit tests were used (Figure 4). Akaike Information Criterion (AIC, 44.52%) was mostly used followed by Deviance (8.45%), Bayesian Information Criterion (BIC, 5.16%) and Pseudo R-Square (4.53%). It was also noted that 4.52% of the papers used AIC and BIC at the same time. Moreover, 27.95% of the selected articles did not report the goodness of fit tests used. It is important to report the goodness of fit tests for assessing whether the test used is the most suitable. For instance, of the selected papers that use AIC as goodness of fit test, more than half should use AICc (corrected AIC) since the sample sizes in those papers were less than 100.

Table 1: Number of selected papers according to journal and research fields

Fields	Journals	n	(%)
Crop production 25 (16.13%)	Field Crop Research	22	(14.19%)
	East African Journal of Sciences	2	(1.30%)
	International Journal of Science	1	(0.65%)
Genetics 5 (3.23%)	Genetic Selection Evolution	2	(1.30%)
	Forest Ecology and Management	1	(0.65%)
	BMC Genetics	2	(1.30%)
Ecology and evolution 27 (17.42%)	BMC Evolution	3	(1.94%)
	Journal of Agricultural Ecology	9	(5.81%)
	Agriculture Ecosystems and Environments	6	(3.87%)
	Ecosystems Services	5	(3.23%)
	Parasite and Vector	4	(2.58%)
Food nutrition 58 (37.42%)	Journal of Agriculture and Food Chemistry	40	(25.80%)
	Food Energy Security	4	(2.58%)
	Journal of Bionutrition Science	1	(0.65%)
	Nutrition Journal	7	(4.52%)
	Food Biotechnology and Nutritional Sciences	6	(3.87%)
	Food Control	39	(25.16%)

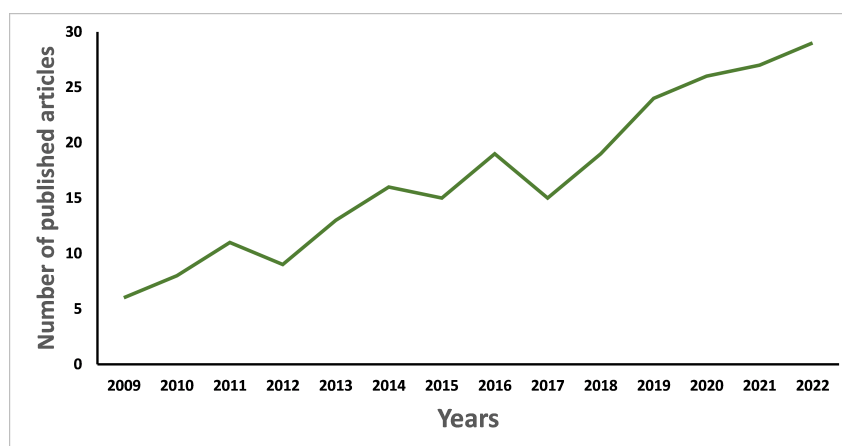


Fig. 2: Trend of the use of GLMs

Furthermore, 59.35% of the selected articles used wald test while 19.35% used likelihood ratio test to assess the significance of the coefficients. In addition, 21.30% of the papers did not report the test used to assess the significance of the coefficients (Table 2). Likelihood ratio tests are not recommended for small and

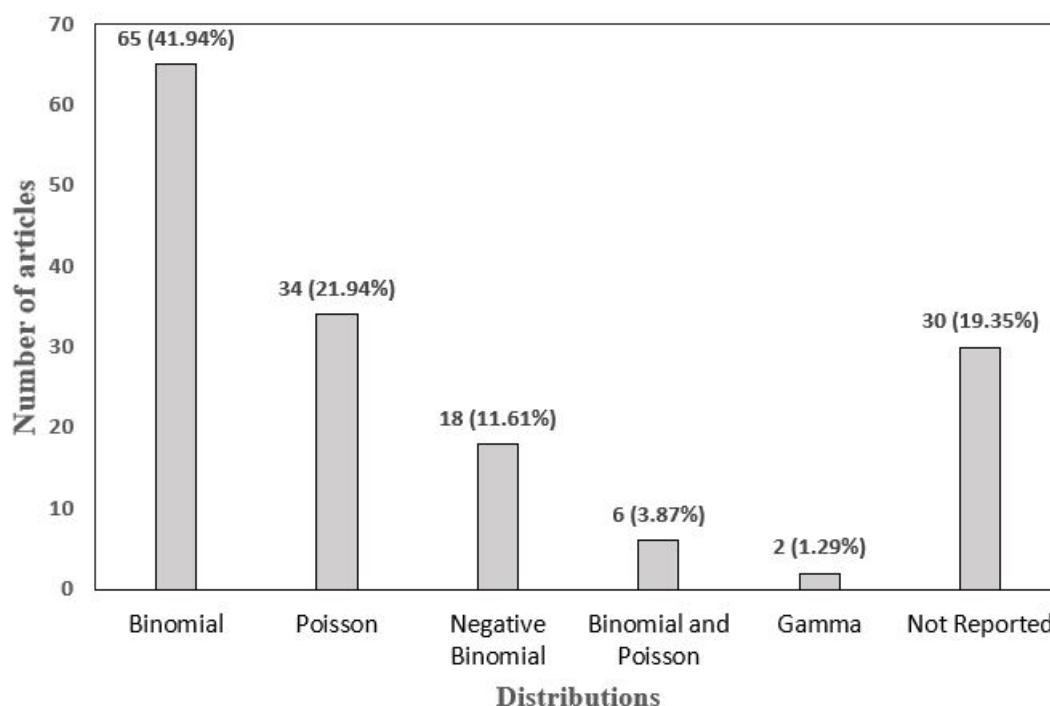


Fig. 3: Distributions of the response variable

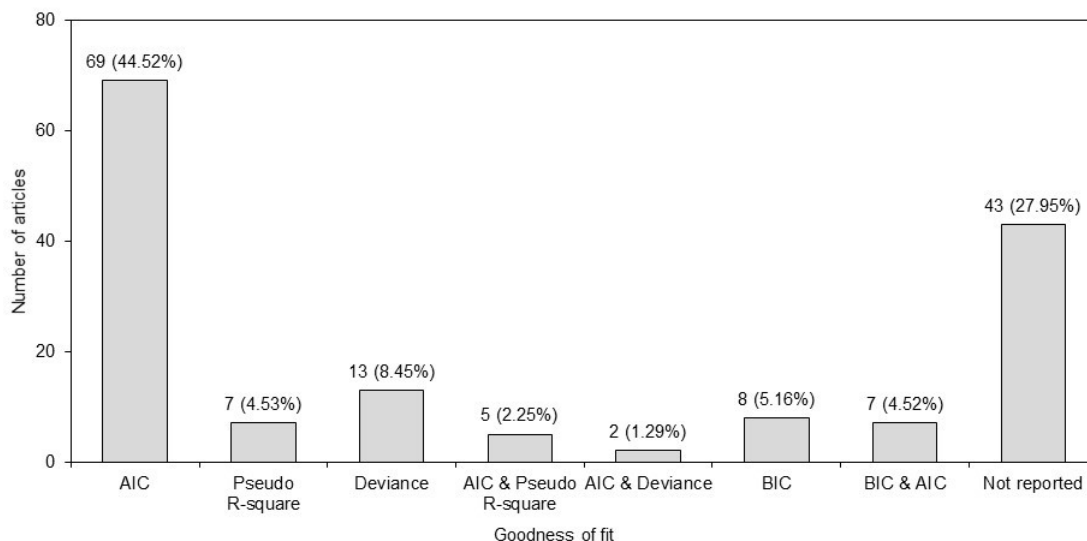


Fig. 4: Goodness of fit tests considered

moderate samples. However, of the selected papers that use likelihood ratio tests for assessing the significance of the coefficients, third had small samples (less than 100).

It was noted that the need to use *GLMs* existed in 93.55% of the papers reviewed. However, in most cases (57.42%), the *GLMs* outputs were not well interpreted (Table 3). The model parameters were interpreted in these papers like in the case of classical linear models. Normally, in *GLMs*, the interpretation of model parameters depends on the link function considered. For instance, for a Poisson model, $\log(\lambda) = \beta_0 + \beta x$, increasing x by 1 unit multiplies the mean value of the response

Table 2: Coefficients significance test

Tests	n	(%)
Likelihood ratio test	30	(19.35%)
Not Reported	33	(21.30%)
Wald Test	92	(59.35%)
Total	155	(100%)

variable by $\exp(\beta)$. In similar terms, for a logistic model, $\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta x$, with p the probability of success, the interpretation of the model parameters is done with odds, $odds = \frac{p}{1-p}$. Indeed, increasing x by 1 unit multiplies the odds by $\exp(\beta)$.

Table 3: Interpretation of GLMs outputs

Good interpretation	n	(%)
Yes	66	(42.58%)
No	89	(57.42%)
Total	155	(100%)

According to the selected articles, the most frequently used software was R (24.51%) followed by SPSS (12.91%) and SAS (11.61%). The remaining articles (50.97%) did not mention the software used.

5. Research and knowledge gaps

While GLMs are increasingly used, there are few studies (Bel *et al.*, 2016; Pekár and Brabec, 2016) guiding users for adequate choice of specific method. From the selected papers, it was observed that most of the outputs (in 57.42% of the papers) were not well interpreted. Moreover, about 21% and 27% did not report the significance of the coefficients and the assessment of the models adequacy, respectively. Therefore, the main consequence is the difficulty to assess the reliability of the results and the validity of conclusions. Collaboration with statisticians may lead to better research designs, data analysis, and interpretation.

The most used GLMs in agriculture and related fields are Poisson, negative binomial, logistic and gamma models (Bourne *et al.*, 2007). One of the advantages of GLMs over ordinary linear models is that GLMs do not require a transformation of the response variable to satisfy application conditions. However, the use of GLMs in agriculture and related fields suffers from some difficulties related to the nature of the data considered.

For count data in agricultural research, the use of the Poisson model suffers from a shortcoming linked to the equi-dispersion assumption (mean equal variance) of the response variable. Indeed, in agriculture and related fields, data often show overdispersion (variance greater than mean). In this situation, the dispersion parameter (variance over mean) is greater than 1. It is often recommended to use negative binomial model as alternative to Poisson model for dispersion parameter greater than 1 (Lokonon, B. and Glèlè Kakaï, 2018). However, there is no threshold of dispersion parameter for which negative binomial model or other alternative models should be considered (Bourne *et al.*, 2007).

In addition, binary data are common in agriculture. For example, presence/absence data are widely used for modelling spatial distribution of the species. Logistic regression is commonly applied to binary data in agriculture

and related fields to predict the distribution of species and vegetation groups under different environmental conditions (Calef *et al.*, 2005). However, this model cannot effectively identify environmental factors responsible for distribution when a species is still expanding its range (Brown, 1994). Indeed, its predictive power over a broad spatial scale is usually low (Guisan *et al.*, 2002). Moreover, GLMs are limited by the fact that they could be used only for fixed factors (McCulloch and Searle, 2001).

In GLMs, parameters are estimated through maximum likelihood estimation (MLE). However, for the overdispersed data, several studies have shown that the use of maximum likelihood estimation is less robust (Nelder and Lee, 1992; Paul and Saha, 2007). A number of alternatives have been discussed in the literature through simulation studies and more robust methods such as extended quasi-likelihood (EQL) have been proposed (Nelder and Pregibon, 1987; Godambe and Thompson, 1989). The primary difficulty of using EQL for applied researchers is that there is currently no software providing a friendly use of this estimation method.

6. General guidelines for an effective use of GLMs

The following steps should be used as guidelines and reported when applying GLMs:

- the first step is to identify the nature and types of variables or factors, and the distribution of the response variable. In the case of non-normal response and only fixed factors in the dataset, the GLMs should be used. Then, the appropriate link function should be identified (for example, Poisson distribution and log link for count data, binomial distribution and logit link for binary data);
- fit the appropriated models;
- assess the model adequacy using for instance the Deviance or Pearson goodness of fit test;
- check the significance of coefficients using Wald test or likelihood ratio test;
- select variables if necessary;
- select the best model if there are several competing models using appropriate selection criteria, such as AIC, BIC, DIC;
- interpret the different coefficients and the relationships between the response and explanatory variables.

These steps are not exhaustive and should be adapted according to the objective of the study. An application guideline in R software is provided in Appendix.

7. Conclusions

The use of GLMs is becoming common in agricultural sciences and related fields due to their flexibility to analyze various data (count, proportional, or binary data). Despite the high level of interest expressed by applied scientists, this study shows that GLMs are not always correctly applied in agricultural sciences. Indeed, in many selected papers, useful and relevant outputs were not reported by users. Close collaboration between statisticians and agricultural scientists is therefore necessary to improve the quality of data analysis and the interpretation of GLMs outputs. Statisticians can provide their expertise in experimental designs, data analysis, and interpretation of parameters while applied scientists can well explain biological phenomena.

Acknowledgment. We are extremely grateful to the editor and the anonymous reviewer for their valuable comments and suggestions, which have Helped to improve the quality of our manuscript.

References

- Bel, L., Daudin, J. J., Etienne, M., Lebarbier, E., Mary-Huard, T., Robin, S. and Vuillet, C. (2016). *Le Modèle Linéaire et ses Extensions*, pp. 1-328.
- Neter, J., Kutner, M. H., Nachtsheim, C. J. and Wasserman, W. (1996). *Applied linear statistical models*. 4th Edition, WCB McGraw-Hill, New York.
- Agresti, A. (2015). *Foundations of linear and generalized linear models*. John Wiley & Sons.
- Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3), 370–384.
- McCullagh, P. and Nelder, J.A. (1989) Generalized Linear Models. 2nd Edition, Chapman and Hall, London. <http://dx.doi.org/10.1007/978-1-4899-3242-6>
- Khuri, A. I., Mukherjee, B., Sinha, B. K., and Ghosh, M. (2006). Design issues for generalized linear models: A review. *Statistical Science*, 21(3), 376–399.
- Hern, A. and Dorn, S. (2001). Statistical modelling of insect behavioural responses in relation to the chemical composition of test extracts. *Physiological Entomology*, 26(4), 381–390.
- Bourne, C., Regular, P., Sun, B., Thompson, S., Trant, A., and Wheeler, J. (2007). Generalized linear model analysis in Ecology, pp. 1-50. Accessed 11 December 2022.
- Pekár, S. and Brabec, M. (2016). *Modern analysis of biological data: generalized linear models in R*. Masarykova univerzita. pp. 1-226. ISBN 978-80-210-8019-5.
- Zuur, A., Ieno, E. N., Walker, N., Saveliev, A. A., and Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R*. Springer Science and Business Media.
- Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., and White, J.S.S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in ecology & evolution*, 24(3), 127–135.
- O'Hara, R. and Kotze, J. (2010). Do not log-transform count data. *Nature Precedings*, 1–1.
- Stroup, W. W. (2015) Rethinking the analysis of non-normal data in plant and soil science. *Agronomy Journal*, 107(2), 811–827.
- Hilbe, J. M. (2011). *Negative binomial regression*. Cambridge University Press.
- Kim, J. S., Kim, D.-K., and Hong, S. J. (2011) Assessment of errors and misused statistics in dental research. *International dental journal*, 61(3), 163–167.
- Guisan, A., Edwards Jr, T. C., and Hastie, T. (2002) Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological modelling*, 157(2-3), 89–100.
- Olsson, U. (2002) *Generalized Linear Models: An Applied Approach*. Student Litteratur Lund., Sweden, 1-232.
- Lokonon, B. and Glèlè Kakaï, R. (2018). Effect of overdispersion and sample size on the performance of Poisson model and its extensions in frame of generalized linear models (glms). *International journal of applied mathematics and statistics*, 57(5), 40–53.
- Calef, M. P., David McGuire, A., Epstein, H. E., Scott Rupp, T., and Shugart, H. H. (2005). Analysis of vegetation distribution in interior alaska and sensitivity to climate change using a logistic regression approach. *Journal of Biogeography*, 32(5), 863–878.
- Brown, D. G. (1994). Predicting vegetation types at treeline using topography and biophysical disturbance variables. *Journal of Vegetation Science*, 5(5), 641–656.
- McCulloch, C. E. and Searle, S. R. (2001). *Generalized, linear, and mixed models (wiley series in probability and statistics)*. Wiley-Interscience
- Nelder, J. and Lee, Y. (1992). Likelihood, quasi-likelihood and pseudolikelihood: some comparisons. *Journal of the Royal Statistical Society: Series B (Methodological)*, 54(1), 273–284.

- Paul, S. and Saha, K. K. (2007). The generalized linear model and extensions: a review and some biological and environmental applications. *Environmetrics: The official journal of the International Environmetrics Society*, 18(4), 421–443.
- Nelder, J. A. and Pregibon, D. (1987). An extended quasi-likelihood function. *Biometrika*, 74(2), 221–232.
- Godambe, V. and Thompson, M. E. (1989). An extension of quasi-likelihood estimation. *Journal of statistical planning and inference*, 22(2), 137–152.
- Ruíz J. S., López O. A. M., Ramírez G. H., Hiriart J. C. (2023) *Generalized Linear Mixed Models with Applications in Agriculture and Biology*. 1st Edition, Springer Cham. <https://doi.org/10.1007/978-3-031-32800-8>

Appendix: Application guideline in R software

Case study

A hunting area manager wanted to model the number of animals killed by hunters per week. He aimed to investigate whether the size of hunter groups and the type of hunting (nocturnal or diurnal) were significantly related to the number of animals killed by hunters per week, and if so, characterize these relationships. We present in the following lines, a step by step procedure to analyze the data.

Step 1: Identify the variables

The variables in the dataset are: *Number_animals* (number of animals killed by hunters); *Hunter_type* (type of hunting (nocturnal or diurnal)) and *Group_size* (number of individuals per hunter group). *Number_animals* is the response variable and *Hunter_type* and *Group_size* are the explanatory variables. The response variable is count discrete and there are no random factor in the explanatory variables. This suggests that GLMs with Poisson model or an extension should be relevant.

Step 2: Load needed libraries and data

```
library(MASS)
library(aod)

Data <- read.table('clipboard', h = T)
attach(Data)
names(Data)

## [1] "Group_size"      "Hunter_type"     "Number_animals"
```

Step 3: compute mean and variance of the response variable as well as dispersion parameter

```
m <- mean(Number_animals)
v <- var(Number_animals)
disp <- v/m
list <- c(m,v,disp)
print(cbind(list))

##           list
## [1,]    5.049505
## [2,]   251.667525
## [3,]    49.840039
```

The mean number of animals is 5.05 and the variance is 251.67 (about 50 times the mean) indicating that the data are overdispersed.

Step 4: fit the models and select the best

Let apply Poisson model as a starting point of a count data model.

```
pois <- glm(Number_animals ~ Group_size + Hunter_type, family='poisson',
data = Data)
#summary(pois)
```

Step 5: assess the model adequacy

Let use Deviance test as goodness of fit test to assess if the model fits well the data. The null hypothesis of this test is H0: the model fits well the data and the alternative hypothesis is H1: the model does not fit well the data.

```
1 - pchisq(summary(pois)$deviance, summary(pois)$df.residual)
## [1] 0
```

P-value related to the test is less than 0.05 meaning that Poisson model does not fit well the data.

As Poisson model does not fit well the data due to the overdispersion, we will consider its extensions. Here, we only consider quasi-Poisson and negative binomial models since there is no zero inflation issues in the data.

We first use Quasi-Poisson model to fit the data. Quasi-Poisson assumes that the variance is proportional rather than equal to the mean.

```
qpois <- glm(Number_animals ~ Group_size + Hunter_type, family = 'quasipoisson',
data = Data)
#summary(qpois)
```

Let use Deviance test to assess if the model fits well the data.

```
1 - pchisq(summary(qpois)$deviance, summary(qpois)$df.residual)
## [1] 0
```

Quasi-Poisson model does not fit well the data ($p < 0.05$).

Let consider negative binomial model. This model is fitted with the function `glm.nb()` of MASS package.

```
nb <- glm.nb(Number_animals ~ Group_size + Hunter_type, data = Data)
#summary(nb)
```

The Deviance test is apply to assess if negative binomial model fits well the data.

```
1 - pchisq(summary(nb)$deviance, summary(nb)$df.residual)
## [1] 0.9040891
```

The P-value is 0.90 (greater than 0.05), then H_0 is accepted and we conclude that negative binomial model fits well the data.

Step 6: assess significance of the coefficients

Let use the Wald test (with the function `wald.test` available in the package `aod`) to assess the significance of the coefficients. The null hypothesis of this test is H_0 : the coefficients are null and the alternative hypothesis is H_1 : At least one of the coefficient is non null.

```
wald.test(b = coef(nb), Sigma = vcov(nb), Terms = 1:2)
## Wald test:
## Chi-squared test:
## X2 = 263.7, df = 2, P(> X2) = 0.0
```

P-value relative to the test is less than 0.05 meaning that the coefficients are significant.

Step 7: interpret the results

The summary of the outputs from the negative binomial model is presented below.


```
summary (nb)

## Call:
## glm.nb(formula = Number_animals ~ Group_size + Hunter_type, data = Data,
##       init.theta = 1.788053289, link = log)
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.73424     0.41339  -4.195 2.73e-05 ***
## Group_size      0.79306     0.09022   8.790 < 2e-16 ***
## Hunter_typeNight -1.85611     0.21953  -8.455 < 2e-16 ***
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Dispersion parameter for Negative Binomial(1.7881) family taken to be 1)
##      Null deviance: 241.559  on 100  degrees of freedom
## Residual deviance:  80.251  on  98  degrees of freedom
## AIC: 444.18
## 2 x log-likelihood:  -436.178
```

On the line corresponding to *Group_size* variable, the coefficient 0.79 associated to this variable means that, one unit increases in the group size multiplies the mean number of killed animals by $\exp(0.79)=2.20$. We can also see that the associated p-value is less than 0.05. We can say that there is a significant association between the two variables.

For the line relating to the hunter type variable, it is noted that the title is *Hunter_typeNight*. Indeed, the results provided on this line concern nocturnal hunting relatively to diurnal hunting who serve as a reference. The results show negative coefficient and significant (p-value ; 0.05) association between nocturnal hunting and number of animals. Then, the mean number of killed animals in nocturnal hunting is lower than those killed in diurnal hunting. In other terms, the mean number of killed animals in nocturnal hunting is $\exp(-1.86)=0.15$ times those killed in diurnal hunting.

The equation of the model is: $\ln(\lambda) = -1.73 + 0.79 \times \text{Group_size} - 1.86 \times \text{Hunter_typeNight}$, with λ the mean number of killed animals.